DOCUMENT RESUME

ED 358 145 TM 019 934

AUTHOR De Ayala, R. J.
TITLE The Influence of Multidimensionality on the Graded Response Model.
PUB DATE Apr 93
NOTE 41p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; Correlation; *Difficulty Level; *Estimation (Mathematics); *Item Bias; Item Response Theory; Mathematical Models; Probability; *Sample Size; *Test Length
IDENTIFIERS Ability Estimates; Dichotomous Variables; *Graded Response Model; Item Discrimination (Tests); *Multidimensionality (Tests); Polytomous Items; Root Mean Square (Statistics); Theta Estimates; Unidimensionality (Tests)

ABSTRACT

Previous work on the effects of dimensionality on parameter estimation was extended from dichotomous models to the polytomous graded response (GR) model. A multidimensional GR model was developed to generate data in one-, two-, and three-dimensions, with two- and three-dimensional conditions varying in their interdimensional associations. Test length (15 and 30 items) and the ratio of sample size to the number of item parameters to estimate were also investigated, using sample sizes of 375 and 750 for the short test and 750 and 1,500 for the longer test. Results show that for unidimensional data a sample size ratio of 5:1 provided reasonably accurate estimation, and that increasing the sample size did not have a significant impact on the accuracy of item parameter estimation. Regardless of data dimensionality, the difficulty parameters were well-estimated, and for the multidimensional data the correlations between estimated item discrimination and the average and the sum of the dimensional discrimination were greater than the correlations between the estimated item discrimination and individual dimensional discriminations. Fidelity coefficients between the mean ability and the ability estimate were greater than those between the ability estimate and the latent traits. The impact of equating on accuracy indices in a multidimensional context was discussed. Seven tables and 16 graphs present analysis data. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

The Influence of Multidimensionality on the Graded Response Model

R.J. De Ayala,
University of Maryland

ED358145

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RALPH DE AYALA
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the National Council on Measurement in Education, April, 1993, Atlanta, GA.

TM019934

ABSTRACT

Most item response theory models assume a unidimensional latent space. This study extended previous work on the effects of dimensionality on parameter estimation from dichotomous models to the polytomous graded response (GR) model. A multidimensional GR model was developed to generate data in one-, two-, and three-dimensions. The two- and three-dimensional conditions contained data sets that varied from one another in their interdimensional association. Moreover, additional factors investigated were test length and the ratio of sample size to the number of item parameters to estimate. Results showed that for the unidimensional data a sample size ratio of 5 : 1 provided reasonably accurate estimation and that increasing the test length from 15 to 30 items did not have a significant impact on the accuracy of item parameter estimation. Regardless of the data's dimensionality, the difficulty parameters were well-estimated and for the multidimensional data the correlations between the estimated item discrimination and the average (as well as the sum of the) dimensional discrimination were greater than the correlations between the estimated item discrimination and the individual dimensional discriminations. Fidelity coefficients between the mean ability and the ability estimate ($\hat{\theta}$) were greater than those between the $\hat{\theta}$ and the latent traits. The impact of equating on accuracy indices in a multidimensional context was discussed.

The Influence of Multidimensionality on the Graded Response Model

To date a number of item response theory (IRT) models have been proposed. One taxonomic scheme for these models is to classify the models as either dichotomous or polytomous (e.g., the Rasch (Rasch, 1980) and Samejima's (1969) graded response (GR) models, respectively). Except for some multidimensional dichotomous models, the majority of IRT models assume a unidimensional latent space. The multidimensional dichotomous models (e.g., McKinley & Reckase, 1983; Sympson, 1978) were developed to overcome the restrictiveness of the unidimensionality assumption and may be classified as either compensatory or noncompensatory. Whereas, Sympson (1978) labeled his model as partially compensatory, however, Way, Ansley, & Forsyth (1988) considered this model to be an example of a noncompensatory multidimensional model. Conceptually, a compensatory model is one in which an examinee's latent traits (θs) interact to produce a response to an item. This interaction may take the form of an examinee's facility on one latent trait (θ_1) compensating for a deficiency in another latent trait (θ_2). In contrast, in a noncompensatory model the examinee's θs do not compensate, per se, for one another to yield a response. Because of difficulties in parameter estimation as well as in the interpretation of the ability space, multidimensional models have yet to obtain widespread acceptance or use in applications. However, it appears that NOHARM (Fraser, 1986) may provide a workable solution to the estimation problem (cf., Miller, 1991). Luecht and Miller (1992) present a unidimensional composite abilities approach for addressing the multidimensionality of some data.

Given that most IRT models assume unidimensionality, several studies (e.g., Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979; Way, Ansley, & Forsyth, 1988) have examined the effect of multidimensionality on unidimensional IRT parameter estimation. These studies have been primarily concerned with the effects of dimensionality on the calibration of a multidimensional data set by either LOGIST (Wingersky, Barton, & Lord, 1982) and/or BILOG (Mislevy & Bock, 1982); both programs are limited to parameter estimation of dichotomous IRT models. Although the models used for data generation differed from one another, the results of these studies have consistently found that multidimensionality affects parameter estimation. In general, when a compensatory multidimensional IRT model was used for data generation, the estimated difficulty ($\hat{b}$) was found to be an estimate of the average of the true difficulties (Way et al., 1988), the estimated discrimination ($\hat{a}$) was an estimate of the sum of the dimensional discriminations (Way et al., 1988), and ability estimates ($\hat{\theta}$) were an estimate of the average true θs (Ackerman, 1989; Way et al., 1988). In contrast, data generation using a noncompensatory model showed that $\hat{b}$ was an overestimate of or correlated more highly with one dimension's difficulty parameters than

with the other dimension's (Ackerman, 1989; Ansley & Forsyth, 1985; Way et al., 1988), $\hat{a}$ was an estimate of the average of the true discriminations (Ansley & Forsyth, 1985; Way et al., 1988), and $\hat{\theta}$ to be an estimate of the average true θs (Ackerman, 1989; Ansley & Forsyth, 1985; Way et al., 1988). In general, these conclusions come from correlational analyses of the estimates with their parameters and an assessment of the accuracy of parameter estimation through the use of the mean absolute difference (a.k.a., MAD or average absolute difference (AAD)). Luecht and Miller (1992) discuss some of the issues associated with ignoring multidimensionality in polytomous data. For instance, they found that item information is reduced when a unidimensional reference composite is fitted to multidimensional polytomous data.

This study's objective was to examine the effect of dimensionality on the parameter estimation of the GR model. Data sets were generated that differed from one another in the number of latent factors as well as their interdimensional association, the number of test items, and the sample size. In this regard, this research extends previous work on the effects of dimensionality on dichotomous model parameter estimation to polytomous models.

METHOD

Model Definition

A multidimensional extension of the GR (MGR) model was developed and used for data generation. This model requires a set of multidimensional θs as well as a set of (multidimensional) item parameters. In the MGR model the examinee responses to item i are categorized into $m_i + 1$ categories, where higher categories indicate greater ability and m_i is the number of category boundaries. Associated with each category of item i is a category score, x_i, with values $0..m_i$. The MGR model may be expressed as:

$$P_{x_i}(\Theta) = \frac{e^{D\sum a_{ih}(\theta_h - d_{x_i})}}{1 + e^{D\sum a_{ih}(\theta_h - d_{x_i})}} \qquad (1),$$

where θ_h is the latent trait on dimension h (h=1..r dimensions), a_{ih} is the discrimination parameter for item i on dimension h, d_{x_i} is the difficulty parameter for category score x for item i, and the summation is across dimensions. A scaling constant, D = 1.702, may be introduced if desired. $P_{x_i}(\Theta)$ is the probability of a randomly selected examinee with latent traits Θ responding in category score x_i or higher for item i; the probability of responding in the lowest category (i.e., P_0) or higher is defined as 1.0 and the probability of responding in the highest category (i.e., P_{m_i+1}) is 0.0. For example, for an item with four response categories (i.e., 0, 1, 2, and 3) $P_2(\Theta)$ is the probability of responding in categories 2 or 3 rather than in categories 0 or 1. Because P_{x_i} is the (cumulative) probability of responding in x_i or higher, the probability of responding in a particular category, $p_{x_i}(\Theta)$, equals the

difference between the cumulative probabilities for adjacent categories (e.g., $p_2(\Theta) = P_2(\Theta) - P_3(\Theta)$). For instance, given m = 3, Θ = {1.00, 1.50}, d = {0.75, 1.250}, a = {1.50, 0.75} and omitting the scaling constant D, one obtains:

$$P_1(\Theta) = \frac{e^{1.50(1.00 - 0.75) + 0.75(1.50 - 0.75)}}{1 + e^{1.50(1.00 - 0.75) + 0.75(1.50 - 0.75)}} = 0.7186$$

$$P_2(\Theta) = \frac{e^{1.50(1.00 - 1.25) + 0.75(1.50 - 1.25)}}{1 + e^{1.50(1.00 - 1.25) + 0.75(1.50 - 1.25)}} = 0.4533$$

Therefore, the probabilities of responding in categories 0, 1, and 2 are:

$$p_0(\Theta) = P_0(\Theta) - P_1(\Theta) = 1.0 - 0.7186 = 0.2814$$

$$p_1(\Theta) = P_1(\Theta) - P_2(\Theta) = 0.7186 - 0.4533 = 0.2653$$

$$p_2(\Theta) = P_2(\Theta) - P_3(\Theta) = 0.4533 - 0.0 = 0.4533$$

When $r \geq 2$ and $m_i = 2$ the MGR reduces to the M2PL (McKinley & Reckase, 1983), if r = 1 the MGR reduces to the GR model, and when r = 1 and $m_i = 2$ (correct and incorrect) the MGR model reduces to the two-parameter model. The option response surfaces (ORS) for the three-step item above are presented in Figures 1a - 1c.

Insert Figures 1a to 1c about here

Design

The data generated differed in terms of the number of latent dimensions and the degree of interdimensional association ($\rho_{\theta_i\theta_j}$), test length, and the ratio of examinees to item parameter estimates. The number of dimensions factor contained three levels: one-, two-, and three-dimensions. The two-factor data contained three degrees of interdimensional associations ($\rho_{\theta_1\theta_2}$ = 0.0, 0.30, 0.75); the first two $\rho_{\theta_1\theta_2}$s were obtained from Ackerman (1989) and the third $\rho_{\theta_1\theta_2}$ was from Wang (1987). The three-dimensional condition contained four data sets that varied from one another in their $\rho_{\theta_i\theta_j}$s ($\rho_{\theta_i\theta_j}$ = 0.0, 0.0, 0.0; $\rho_{\theta_i\theta_j}$ = 0.30, 0.30, 0.30; $\rho_{\theta_i\theta_j}$ = 0.30, 0.30, 0.75; $\rho_{\theta_i\theta_j}$ = 0.75, 0.75, 0.75).

The test length factor contained two levels, 15 and 30 items, where the 15 items were randomly selected from the 30-item test. The sample size ratio factor consisted of two ratios of examinees to item parameter estimates, 5 to 1 and 10 to 1. These two ratios resulted in sample sizes of 375 and 750 for the 15-item test and 750 and 1500 for the 30-item test.

Therefore, the study's design consisted of sample size ratio by test length by dimensionality (2 X 2 X 8 = 32 cells). For each cell 15 replications were generated and all of the 480 (=15 x 32) data sets were unique. For each data set item parameter estimates for the GR model were obtained using MULTILOG 5.1 (Thissen, 1988).

Data

For the unidimensional data set θs were randomly sampled from a unit normal distribution. The two- and three-dimensional conditions were created by randomly sampling θs from a multinormal distribution with known $\rho_{\theta_i\theta_j}$. For each data set the appropriate number of zs were randomly sampled from the relevant distribution and their responses to 5-choice items were generated; the zs were taken to be the simulees' θ(s). In the following and unless otherwise noted, the subscript on the discrimination parameters refers to the dimension and the subscript for the item, i, will be omitted. The d_Xs used in the response string generation were identical to the b_Xs used in Dodd, Koch, and De Ayala (1989). Dodd, Koch, and De Ayala generated their b_Xs so that they would distribute the items uniformly across the θ continuum (as expressed by their category boundaries) while at the same time representing values obtained from real data. The a_hs were randomly sampled from a uniform distribution [0.80, 2.0]. For the unidimensional data the a_1s were used in generating the response data and for the bidimensional data the a_1s and a_2s were used.

For each data set the θs plus the relevant item parameters were used to generate polytomous response strings with a random error component for each simulated examinee. For the multidimensional and unidimensional data sets the generation of an examinee's polytomous response string was accomplished by calculating the probability of responding to each item alternative according to the MGR model; the scaling factor D was set to 1.0. Based on the probability for each alternative, cumulative probabilities were obtained for each alternative. A random error component was incorporated into each response by selecting a random number from a uniform distribution [0, 1] and comparing it to the cumulative probabilities. The ordinal position of the first cumulative probability that was greater than the random number was taken as the examinee's response to the item.

Equating

The Stocking and Lord (1983) procedure, as implemented in Equate (Baker, Al-Karni, & Al-Dosary, 1992), was used to place the item parameter estimates on the same scale as their parameters. The equating was done at 21 theta points; Baker (1992) contains a discussion of the procedure used.

Analyses

Descriptive statistics and Pearson product-moment correlation coefficients between $\hat{a}$ and the a_h(s), the average of the a_hs across dimensions ($\bar{a}$), and the sum of the dimensional a_hs ($\sum a$) as well as between $\hat{b}_{x_i}$ and d_{x_i} were computed for each replication and averaged across replications. Analysis of the accuracy of the item parameter estimation involved calculating root mean square error (RMSE), and Bias. RMSE and Bias were calculated according to:

$$RMSE(\Phi) = \sqrt{\frac{\sum (x - \Phi)^2}{n}}, \quad (2)$$

$$Bias(\Phi) = \frac{\sum (x - \Phi)}{n}, \quad (3)$$

where x was either $\hat{b}_{x_i}$ (i.e., the difficulty estimate for category x of item i) or $\hat{a}_i$ (the discrimination parameter estimate of item i), and n was the number of replications. For item difficulty Φ was d_{x_i} and for item discrimination RMSE(Φ) and Bias(Φ) were calculated with respect to each a_h, the $\bar{a}$, and the $\sum a$ (i.e., Φ was a_h, or $\bar{a}$, or $\sum a$). The accuracy of the item parameter estimates for the 15-item test were compared to the estimates of the same items embedded in the 30-item test. RMSE and Bias were treated as the dependent variables in a one-group repeated measure design to determine whether they were significantly affected by test length (within subjects) and the sample size ratio (between subjects); the Bonferroni procedure was used to control for experimentwise Type I error rate.

For ability, Φ was set to the the true ability and x was the $\hat{\theta}$. Correlations (fidelity coefficients) were calculated between the $\hat{\theta}$s and the θ_hs as well as between $\hat{\theta}$ and $\bar{\theta}$ ($r_{\hat{\theta}\theta_h}$ and $r_{\hat{\theta}\bar{\theta}}$, respectively). The correlations were calculated for each replication and averaged across replications.

Because the true abilities were randomly generated each examinee had potentially unique true abilities. Therefore, for ability RMSE and Bias were calculated in two ways: (a) across all examinees for each replication and across replications, and (b) across replications but as a function of ability. In this latter case, it was necessary to group the examinees so that the calculations of RMSE and Bias were based on more than one examinee at each theta point. Therefore, the true abilities were rounded to one decimal place and the examinees having the same rounded true ability were used for calculating RMSE at that particular theta point.

RESULTS

Component analyses of the covariance matrix for each (multidimensional) level of the number of dimensions factor showed that the $\rho_{\theta_i\theta_j}$ = 0.0, 0.0, 0.0 level contained three factors each accounting for 33.3% of the total variance (σ^2_{total}), the $\rho_{\theta_i\theta_j}$ = 0.30, 0.30, 0.30 level contained a dominant first factor and two additional factors each of which accounted for 23.3% of σ^2_{total}, the $\rho_{\theta_i\theta_j}$ = 0.30, 0.30, 0.75 level's distribution of σ^2_{total} across the three factors was 64.71%, 26.96%, and 8.33%, and the fourth level ($\rho_{\theta_i\theta_j}$ = 0.75, 0.75, 0.75) contained a single factor accounting for 83.3% of the σ^2_{total} with the remaining factors each accounting for 8.3% of the total variability. Therefore, the component analyses appear to support the fact that the data possessed the intended characteristics.

Tables 1 and 2 contain the average Pearson correlation coefficients (across replications) between the item parameters and their estimates. For the unidimensional data sets the correlations between $\hat{a}$ and a_1 increased as the sample size ratio increased for a given test length (Table 1). Moreover, for the unidimensional data and for a given sample size ratio the correlations were higher for the 30-item test than for the 15-item test. This increase in the correlation was not due to the $\hat{a}$s for the 30-item test having greater variability than those of the 15-item test. (For the 5 : 1 sample size ratio the standard deviation (s) for the $\hat{a}$s based on the 30-item test was 0.355 and for the 15-item test $s_{\hat{a}} = 0.435$, whereas for the 10 : 1 sample size ratio for the 30- and 15-item tests the $s_{\hat{a}} = 0.355$ and $s_{\hat{a}} = 0.403$, respectively; the s of the as for the 30-item test was 0.360 and for the 15-item test it was 0.335.)

Insert Table 1 about here

Except for the 15-item test data sets (sample size ratio 5 : 1), as the data became progressively more unidimensional the correlations between $\hat{a}$ and the a_hs, as well as between $\hat{a}$ and $\bar{a}$, increased. The addition of a third factor led to a decrease in the $r_{\hat{a}\bar{a}}$s and $r_{\hat{a}a_h}$. In addition, the $r_{\hat{a}\bar{a}}$ for the bidimensional $\rho_{\theta_1\theta_2} = 0.0$ level was larger than that for the tridimensional level (all $\rho_{\theta_i\theta_j} = 0.75$). Comparisons of the $r_{\hat{a}\bar{a}}$s to the $r_{\hat{a}a_h}$s for the multidimensional data sets showed that, in general, $\hat{a}$ had a stronger linear relationship with $\bar{a}$ and $\sum a$ than with the individual a_hs.

Table 2 shows that, in general, the $\hat{b}_x$s were highly linearly related to their corresponding d_xs. As can be seen, $\hat{b}_1$ tended to be more highly related to d_1 than were the $\hat{b}_x$s and d_xs for the other category boundaries. Furthermore, as one progressed from the second to the fourth category boundary the $r_{\hat{b}_x d_x}$ decreased. This was true regardless of the dimensionality of the data. In general, as the two- and three-dimensional data sets became more unidimensional the $r_{\hat{b}_x d_x}$s increased and the $r_{\hat{b}_x d_x}$s based on the multidimensional data were higher than were the corresponding category $r_{\hat{b}_x d_x}$s based on the unidimensional data. This pattern of $r_{\hat{b}_x d_x}$s was associated with standard deviations for the $\hat{b}_x$s based on the multidimensional data that were larger than the standard deviations for the $\hat{b}_x$s based on the unidimensional data. In general, for a given sample size ratio, the $r_{\hat{b}_x d_x}$s tended to be higher for the 30-item test than for the 15-item test and the $r_{\hat{b}_x d_x}$s tended to be larger for the 10 : 1 ratio than for the 5 : 1 ratio. In addition, regardless of the data's dimensionality, the test length, and the sample size ratio, there was a tendency for the standard deviations of the $\hat{b}_x$s to increase as one progressed from category 1 to category 5. For instance, for the 15-item test/10 : 1 sample size ratio/unidimensional data the standard deviations for b_1, b_2, b_3 and b_4 were 0.64, 1.09, 1.19, and 1.39, respectively, and for

the 30-item test/10 : 1 sample size ratio/bidimensional ($\rho_{\theta_1\theta_2}$= 0.0) data the standard deviations were $s\hat{b}_1$ = 1.28, $s\hat{b}_2$ = 1.45, $s\hat{b}_3$ = 1.56, and $s\hat{b}_4$ = 2.07.

Insert Table 2 about here

Table 3 contains the Summary Tables for the analysis of the RMSE(a) and Bias(a) for the unidimensional data. As can be seen, neither the sample size ratio nor the test length had a significant effect on the accuracy of estimation. Figure 2 contains the corresponding RMSE and Bias plots. The RMSE plot reflects the finding that test length and sample size ratio did not have an effect on RMSE(a). Moreover, MULTILOG exhibited a slight reduction in the accuracy of estimation as a increased; this inaccuracy was due to an increase in overestimating a.

Insert Table 3 and Figure 2 about here

Analysis of the difficulty parameters (Table 4) showed that there was a significant test length by sample size ratio interaction only in the estimation accuracy of d_1. Post hoc analyses showed that for the 5 : 1 sample size ratio the accuracy in estimating d_1 based on the 15-item test was significantly greater than that based on the 30-item test. Moreover, for the 15-item test the RMSE(d_1) increased when the sample size ratio was doubled. Similarly, the bias analysis showed that the Bias(d_1) for 15-item test/5 : 1 sample size ratio was significantly less than that for either the 30-item test/5 : 1 sample size ratio or the 15-item test/10 : 1 sample size ratio. There were no statistically significant findings for d_2, d_3, or d_4.

Insert Table 4 about here

RMSE(d_x) plots for the unidimensional data are presented in Figure 3. As can be seen, for the 15-item test/5 : 1 sample size ratio d_1 was comparatively well-estimated (Figure 3a), but that for d_2, d_3 and d_4 this condition yields less accurate estimates (Figure 3b to Figure 3d, respectively) than the other conditions. There appeared to be a tendency for a decrease in the accuracy of estimation of d_2, d_3 and d_4 as these difficulty parameters became more difficult (e.g., d_4 = 2.0).

Insert Figure 3 about here

Figure 4 presents the Bias(d_x) plots. As was the case with the RMSE(d_x) plots, for the 5 : 1 sample size ratio/15-item test there was less bias in estimating d_1 than in estimating d_1 under the other conditions. This pattern was reversed for d_2, d_3 and d_4

and, in general, there appeared to be a tendency for an increase in the overestimation bias for d_2, d_3 and d_4 as these difficulty parameters became more difficult.

Insert Figure 4 about here

The average fidelity coefficients across replications are presented in Table 5. For a given dimensionality the fidelity coefficients based on the 30-item test were greater than those for the 15-item test regardless of the sample-size ratio. Overall, the $r_{\hat{\theta}\bar{\theta}}$s were greater than the $r_{\hat{\theta}\theta_h}$, regardless of the data's dimensionality. For a given test length/sample size ratio the $r_{\hat{\theta}\bar{\theta}}$s were higher with the multidimensional data than they were with the unidimensional data.

Insert Table 5 about here

Table 6 contains the average RMSE and Bias for ability across examinees and replications. As can be seen, the mean RMSEs for the unidimensional data are comparable to those found by Reise and Yu (1990). Increasing the test length resulted in a reduction in the average RMSEs, however, increasing the sample size ratio led to an increase in the average RMSE. In general, there appears to be very little overall bias in estimating θ, although there is a slight tendency to underestimate. These averages are potentially misleading. Figure 5 contains RMSE and Bias plots for the estimation of θ. As can be seen, the RMSE(θ) was relatively consistent regardless of the sample size ratio or the test length. The Bias plot showed that there was only a slight underestimation bias around $-1.5 \leq \theta \leq 0.75$, although the 15-item test tended to result in less Bias across the θ scale than did the 30-item test. It should be noted that RMSE and Bias values outside the -2.0 to 2.0 ability range are based on relatively small numbers of examinees sizes, and therefore, are less stable and should have little significance attached to them.

Insert Table 6 and Figure 5 about here

DISCUSSION

The number of alternatives was not a factor in this study. However, the present results in conjunction with those of Ackerman (1989) using two category items appear to indicate that the general findings should not be influenced by the number of item alternatives.

Reise and Yu (1990) have recommended that a minimum of 500 examinees should be used to obtain accurate and stable estimates of the unidimensional GR item parameters. However, we feel that in general such guidelines are more useful if stated in terms of the ratio

of examinees to item parameters to be estimated. For instance, in this study comparatively reasonably accurate RMSEs were achieved with 375 examinees. That is, it appears that a ratio of examinees to item parameter estimates of 5 : 1 provides reasonable item parameter estimation. This 5 : 1 ratio is consistent with the Reise and Yu (1990) suggestion of the use of 500 examinees because their study used 25 four-choice items. However, it should be noted that regardless of the sample size, with polytomous models it is the distribution of responses across the item alternatives that will result in accurate and stable item/category parameter estimation. As an extreme example, consider a 10-item test (4 option items) administered to 40,000 examinees (ratio of 10,000 : 1). If all of the examinees respond only in the first category, the item parameters for the other categories will be "poorly" estimated. With larger sample sizes this problem is less likely to occur.

For the purposes of the study the replication samples could have been assumed to be randomly equivalent. Because the coefficients from the equating of each replication to the parameter scale were similar to one another the assumption that the replications were more or less equivalent would be confirmed. However, strictly speaking simply because the replications were essentially equivalent to one another does not imply that the estimates' scale will be the same as the parameter scale. For instance, Table 7 contains the repeated measures analysis for a when the item parameter estimates were not equated to the parameter scale. As can be seen, with the unequated estimates there was a significant test length main effect; doubling the 15-item test produced a significant decrease in RMSE(a) from 0.392 for the 15-item test to 0.168 for the 30-item test. However, this effect is an artifact attributable to the use of a scale dependent accuracy index with noncomparable scales. Figure 6 contains the corresponding RMSE(a) plot depicting the effect of test length. Moreover, a comparison of Figures 6 and 2 shows that when the item estimates are not equated to the parameter scale, the estimates appear to be more accurate when they are *not* equated than when they *are* equated and that the order of the conditions' RMSEs conditional on θ is not the same across figures (e.g., compare the two figures' RMSEs at a = 1.1 and a = 1.6). Because at present there is no way to equate the unidimensional parameter estimates to the multidimensional parameter scale, the use of scale dependent accuracy indices, such as RMSE, Bias, MAD (or AAD), for assessing the effect of multidimensionality on the accuracy of estimation is ill-advised and inappropriate.

Insert Table 7 and Figure 6 about here

The use of correlations for the assessment of the (linear) relationship between estimates and parameters may be used. In this regard, the influence of multidimensionality

on parameter estimation was reflected in an overall decrease in $r_{\hat{a}\bar{a}}$ and $r_{\hat{a}a_h}$ as the number of factors in the data increased and as the interdimensional correlation decreased. For the multidimensional data, $\hat{a}$ had a stronger linear relationship with $\bar{a}$ than with the individual a_hs. However, because $r_{\hat{a}\bar{a}} = r_{\hat{a}\sum a}$ there is no way to determine whether $\hat{a}$ was an estimate of the sum of the dimensional discriminations or the average dimensional discrimination; the equating issue discussed above negates the use of accuracy indices for deciding between $\sum a$ and $\bar{a}$. Furthermore, the poorer accuracy indices others have found in multidimensional situations may be more a function of the large values that may arise with $\sum a$ than anything intrinsic to the taken of a sum. For example, large RMSE values for $\sum a$ may be primarily a result of the fact that the data simply do not reflect items that discriminate to a degree characterized by $\sum a$ (e.g., the data were generated with $0.8 \leq a \leq 2.0$ and $\sum a > 2.5$). The fact that the transformation is a sum of a_h as oppose to the taking of an average of a_h may be irrelevant. What is more important is that the value of the transformation (either a sum or an average) fall within a range represented by the data (e.g., 0.8 to 2.0). Conceptually then, for $\sum a$s that are comparable in magnitude to $\bar{a}$s the corresponding accuracy indices for $\sum a$ and $\bar{a}$ should be similar to one another.

In addition to the equating issue there is an additional problem concerned with rotational indeterminacy. That is, the latent ability space does not have a unique orientation and the dimensions may be rotated without affecting $P_{x_i}(\boldsymbol{\Theta})$ or $p_{x_i}(\boldsymbol{\Theta})$. Therefore, different $\boldsymbol{\Theta}$ and $\boldsymbol{a}$ will produce identical $P_{x_i}(\boldsymbol{\Theta})$, $p_{x_i}(\boldsymbol{\Theta})$, and option response surfaces. For instance, if one rotates the axes 90° the transformed abilities become $\boldsymbol{\Theta}' = \{-1.50, 1.00\}$ and the correspondingly transformed discriminations $\boldsymbol{a}' = \{-0.15, 2.40\}$. Omitting the scaling constant D and letting $\boldsymbol{d} = \{0.75, 1.250\}$ one obtains:

$$P_1(\boldsymbol{\Theta}) = \frac{e^{-0.15(-1.50 - 0.75) + 2.40(1.00 - 0.75)}}{1 + e^{-0.15(-1.50 - 0.75) + 2.40(1.00 - 0.75)}} = 0.7186$$

$$P_2(\boldsymbol{\Theta}) = \frac{e^{-0.15(-1.50 - 1.25) + 2.40(1.00 - 1.25)}}{1 + e^{-0.15(-1.50 - 1.25) + 2.40(1.00 - 1.25)}} = 0.4533 ,$$

and the probabilities of responding in categories 0, 1, and 2 are:

$$p_0(\boldsymbol{\Theta}) = P_0(\boldsymbol{\Theta}) - P_1(\boldsymbol{\Theta}) = 1.0 - 0.7186 = 0.2814$$

$$p_1(\boldsymbol{\Theta}) = P_1(\boldsymbol{\Theta}) - P_2(\boldsymbol{\Theta}) = 0.7186 - 0.4533 = 0.2653$$

$$p_2(\boldsymbol{\Theta}) = P_2(\boldsymbol{\Theta}) - P_3(\boldsymbol{\Theta}) = 0.4533 - 0.0 = 0.4533$$

These are the same $P_{x_i}(\boldsymbol{\Theta})$ and $p_{x_i}(\boldsymbol{\Theta})$ obtained above with $\boldsymbol{\Theta} = \{1.00, 1.50\}$, $\boldsymbol{a} = \{1.50, 0.75\}$, and $\boldsymbol{d} = \{0.75, 1.250\}$. Clearly, this indeterminacy may also affect the assessment of estimation accuracy.

While Hirsch (1989) has explored the equating of multidimensional models to one another, his results were not completely satisfactory and more research needs to be concern with equating with multidimensional models. First, because comparison studies such as this one and the others discussed above require equating and, second, because if multidimensional models are to become a viable approach to measurement, then horizontal and vertical equating issues will need to be addressed. In this regard, Wang's (1987) reference composite may provide a pragmatic approach to this problem when the latent traits are linearly independent.

REFERENCES

Ackerman, T.A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Baker, F. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Baker, F., Al-Karni, A., & Al-Dosary, I.M. (1992). *Equate* (Version 2.0) [Computer Program]. Laboratory of Experimental Design, Department of Educational Psychology, University of Wisconsin, Madison, WI.

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-144.

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-200.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [Computer Program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Hirsch, T.M. (1989). Multidimensional Equating. *Journal of Educational Measurement, 26*, 337-349.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Luecht, R.M., & Miller, T.R. (1992). Unidimensional calibrations and interpretations of composite abilities for multidimensional tests. *Applied Psychological Measurement, 16*, 279-294.

Luecht, R.M., & Miller, T.R. (1992, April). *Considerations of multidimensionality in polytomous item response models.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

McKinley, R., & Reckase, M.D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR83-2). Iowa City, IA: American College Testing Program.

Miller, T.R. (1991). *Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program* (Research Report ACT 91-2). Iowa City, IA: American College Testing Program.

Mislevy, R.J., & Bock, R.D. (1982). *BILOG, maximum likelihood item analysis and test scoring: Logistic model.* Mooresville, IN: Scientific Software, Inc.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960)

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, (No. 17).

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp 82-98). Minneapolis: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Thissen, D.J. (1988). *MULTILOG-User's Guide* (Version 5.1). Scientific Software, Inc. Mooresville, IN.

Wang, M. (1987, April). *Estimation of ability parameters from response data to items that are precalibrated with a unidimensional model.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

Acknowledgements

The author gratefully acknowledges Dr. F. Baker for making his Equate program available, and the helpful comments of two anonymous reviewers, Drs. M. Reckase, R. Leucht, and W.D. Schafer.

Table 1

Average correlations between $\hat{a}$ and a_1, a_2, a_3, $\bar{a}$.

Sample size ratio	Test Length	Parameter	I[a]	II[b]	II[c]	II[d]	III[e]	III[f]	III[g]	III[h]
5 : 1	15	a_1	0.920	0.662	0.684	0.689	0.669	0.655	0.606	0.700
		a_2		0.765	0.752	0.744	0.676	0.696	0.704	0.593
		a_3					0.530	0.542	0.574	0.592
		$\bar{a}$[i]		0.910	0.914	0.911	0.872	0.882	0.882	0.873
	30	a_1	0.954	0.491	0.577	0.654	0.488	0.563	0.493	0.608
		a_2		0.782	0.729	0.691	0.605	0.590	0.628	0.548
		a_3					0.246	0.254	0.311	0.305
		$\bar{a}$[i]		0.895	0.915	0.939	0.813	0.852	0.866	0.880
10 : 1	15	a_1	0.946	0.689	0.714	0.768	0.675	0.717	0.644	0.673
		a_2		0.781	0.776	0.735	0.712	0.693	0.706	0.673
		a_3					0.540	0.557	0.645	0.652
		$\bar{a}$[i]		0.937	0.948	0.950	0.897	0.913	0.932	0.932
	30	a_1	0.959	0.505	0.563	0.613	0.568	0.589	0.530	0.630
		a_2		0.780	0.771	0.739	0.583	0.627	0.646	0.564
		a_3					0.201	0.238	0.307	0.302
		$\bar{a}$[i]		0.903	0.936	0.946	0.821	0.882	0.897	0.901

Notes: [a]unidimensional, [b]$\rho_{\theta_1\theta_2} = 0.0$, [c]$\rho_{\theta_1\theta_2} = 0.30$, [d]$\rho_{\theta_1\theta_2} = 0.75$, [e]$\rho_{\theta_i\theta_j} = 0.0, 0.0, 0.0$, [f]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.30$, [g]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.75$, [h]$\rho_{\theta_i\theta_j} = 0.75, 0.75, 0.75$, [i]These average correlations between $\hat{a}$ and $\bar{a}$ are the same as would be obtained between the $\hat{a}$ and $\sum a$

Table 2

Average correlations between $\hat{b}$ and d_1, d_2, d_3, d_4.

Sample size ratio	Test Length	Parameter	I[a]	II[b]	II[c]	II[d]	III[e]	III[f]	III[g]	III[h]
5 : 1	15	d_1	0.987	0.993	0.992	0.994	0.967	0.983	0.993	0.995
		d_2	0.954	0.983	0.986	0.986	0.992	0.994	0.995	0.993
		d_3	0.907	0.959	0.967	0.972	0.956	0.990	0.992	0.991
		d_4	0.913	0.847	0.950	0.978	0.871	0.949	0.988	0.990
	30	d_1	0.992	0.995	0.996	0.997	0.990	0.998	0.997	0.998
		d_2	0.931	0.981	0.983	0.987	0.994	0.995	0.995	0.995
		d_3	0.911	0.967	0.976	0.982	0.991	0.993	0.993	0.994
		d_4	0.891	0.935	0.974	0.979	0.945	0.977	0.993	0.994
10 : 1	15	d_1	0.990	0.996	0.996	0.997	0.981	0.996	0.997	0.998
		d_2	0.959	0.984	0.987	0.988	0.994	0.995	0.995	0.994
		d_3	0.913	0.965	0.973	0.975	0.983	0.988	0.991	0.992
		d_4	0.913	0.953	0.973	0.975	0.921	0.976	0.989	0.991
	30	d_1	0.995	0.997	0.998	0.998	0.993	0.998	0.999	0.998
		d_2	0.934	0.981	0.984	0.987	0.988	0.995	0.995	0.995
		d_3	0.915	0.971	0.975	0.981	0.984	0.994	0.994	0.995
		d_4	0.897	0.954	0.971	0.977	0.944	0.993	0.994	0.995

Notes: [a]unidimensional, [b]$\rho_{\theta_1\theta_2} = 0.0$, [c]$\rho_{\theta_1\theta_2} = 0.30$, [d]$\rho_{\theta_1\theta_2} = 0.75$, [e]$\rho_{\theta_i\theta_j} = 0.0, 0.0, 0.0$, [f]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.30$, [g]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.75$, [h]$\rho_{\theta_i\theta_j} = 0.75, 0.75, 0.75$

Table 3

Summary Table for unidimensional data set: RMSE(a)

Source	SS	df	MS	F	p
Between					
Ratio[a]	0.0029	1	0.0029	0.115	0.736
Items w/i Ratio[a]	1.2400	48	0.0258		
Within					
Test Length	0.0211	1	0.0211	1.381	0.274
Ratio x Test Length	0.0011	1	0.0011	0.072	0.795
Error	0.1222	8	0.0153		

Summary Table for unidimensional data set: Bias(a)

Source	SS	df	MS	F	p
Between					
Ratio[a]	0.0003	1	0.0003	0.010	0.919
Items w/i Ratio[a]	1.2584	48	0.0262		
Within					
Test Length	0.0140	1	0.0140	1.008	0.345
Ratio x Test Length	0.0003	1	0.0003	0.022	0.887
Error	0.1111	8	0.0139		

Note: Ratio[a]: sample size ratio

Table 4

Summary of RMSE(d_1) analysis for unidimensional data

Source	SS	df	MS	F	p
Between					
Ratio[a]	0.3185	1	0.3185	9.858	0.003*
Items w/i Ratio[a]	1.5807	48	0.0329		
Within					
Test Length	0.0767	1	0.0767	3.847	0.086
Ratio x Test Length	0.1638	1	0.1638	8.218	0.021*
Error	0.1594	8	0.0199		

RMSE Cell Means: Ratio[a] x Test Length

	Test Length	
Ratio[a]	15	30
5 : 1	0.182	0.450
10 : 1	0.499	0.445

Summary of Bias(d_1) analysis for unidimensional data

Source	SS	df	MS	F	p
Between					
Ratio[a]	0.6413	1	0.6413	17.340	0.000*
Items w/i Ratio[a]	1.8157	48	0.0378		
Within					
Test Length	0.1371	1	0.1371	6.833	0.031
Ratio x Test Length	0.2400	1	0.2400	11.962	0.009*
Error	0.1605	8	0.0201		

Bias Cell Means: Ratio[a] x Test Length

	Test Length	
Ratio[a]	15	30
5 : 1	-0.060	-0.443
10 : 1	-0.493	-0.441

Note: Ratio[a]: sample size ratio; *significant at overall $\alpha = 0.05$

Table 5

Average fidelity coefficients

Sample size ratio	# of items	Theta	I[a]	II[b]	II[c]	II[d]	III[e]	III[f]	III[g]	III[h]	N
5 : 1	15	θ_1	0.948	0.659	0.773	0.899	0.550	0.693	0.641	0.884	5625
		θ_2		0.709	0.796	0.913	0.552	0.725	0.842	0.894	5625
		θ_3					0.586	0.725	0.844	0.893	5625
		$\bar{\theta}$		0.963	0.968	0.969	0.967	0.974	0.975	0.977	5625
	30	θ_1	0.963	0.677	0.783	0.912	0.553	0.699	0.644	0.896	11,250
		θ_2		0.692	0.792	0.915	0.542	0.710	0.841	0.896	11,250
		θ_3					0.593	0.728	0.850	0.900	11,250
		$\bar{\theta}$		0.974	0.975	0.977	0.977	0.980	0.981	0.983	11,250
10 : 1	15	θ_1	0.948	0.657	0.769	0.904	0.535	0.700	0.634	0.888	11,250
		θ_2		0.703	0.790	0.913	0.572	0.721	0.844	0.887	11,250
		θ_3					0.569	0.718	0.845	0.894	11,250
		$\bar{\theta}$		0.962	0.968	0.970	0.967	0.973	0.975	0.977	11,250
	30	θ_1	0.963	0.675	0.780	0.914	0.550	0.708	0.647	0.893	22,500
		θ_2		0.697	0.791	0.914	0.554	0.709	0.847	0.894	22,500
		θ_3					0.585	0.732	0.855	0.901	22,500
		$\bar{\theta}$		0.974	0.976	0.978	0.977	0.980	0.982	0.983	22,500

Notes: [a]unidimensional, [b]$\rho_{\theta_1\theta_2} = 0.0$, [c]$\rho_{\theta_1\theta_2} = 0.30$, [d]$\rho_{\theta_1\theta_2} = 0.75$, [e]$\rho_{\theta_i\theta_j} = 0.0, 0.0, 0.0$, [f]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.30$, [g]$\rho_{\theta_i\theta_j} = 0.30, 0.30, 0.75$, [h]$\rho_{\theta_i\theta_j} = 0.75, 0.75, 0.75$

Table 6

Average RMSE/Bias for ability (unidimensional data)

Sample size ratio	# of items	RMSE	Bias	N
5 : 1	15	0.381	-0.130	5625
	30	0.353	-0.058	11,250
10 : 1	15	0.444	-0.140	11,250
	30	0.435	-0.127	22,500

Table 7

Summary Table for unidimensional data set: RMSE(a) - unequated parameters

Source	SS	df	MS	F	p
Between					
Ratio[a]	0.0107	1	0.0107	1.562	0.217
Items w/i Ratio[a]	0.3314	48	0.0069		
Within					
Test Length	0.3188	1	0.3188	62.301	0.000*
Ratio x Test Length	0.0022	1	0.0022	0.438	0.527
Error	0.0409	8	0.0051		

Note: Ratio[a]: sample size ratio; *significant at overall $\alpha = 0.05$

Figure Captions

Figure 1. Option response surfaces for a three-category item with $\boldsymbol{d}$ = {0.75, 1.250} and $\boldsymbol{a}$ = {1.50, 0.75}
Figure 1a: ORS for category 1
Figure 1b: ORS for category 2
Figure 1c: ORS for category 3

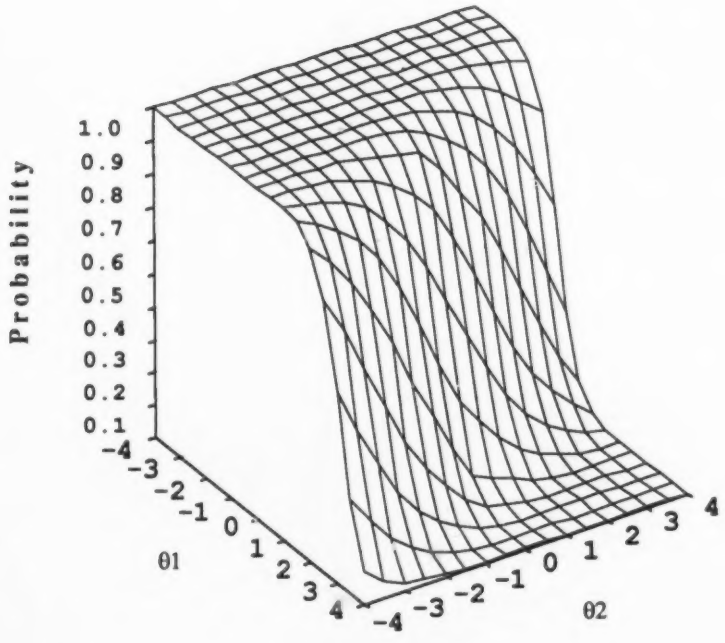
Figure 2. RMSE(a) and Bias(a) for unidimensional data
Figure 2a: RMSE(a)
Figure 2b: Bias(a)

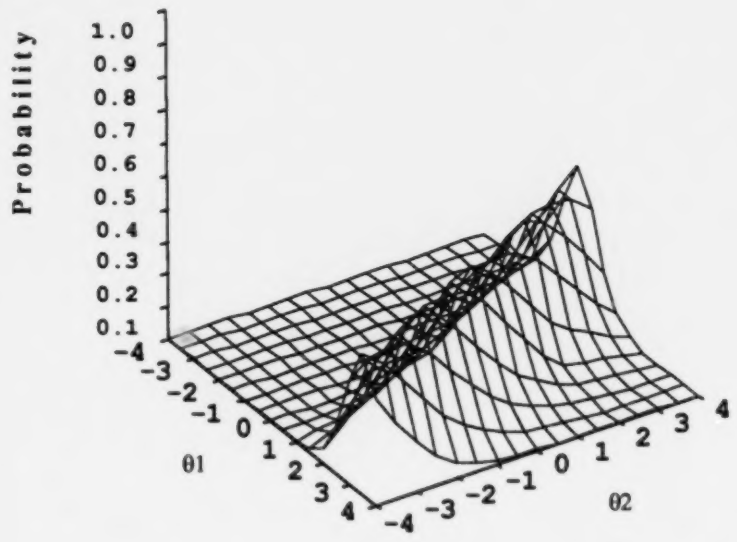
Figure 3. RMSE(d_1), RMSE(d_2), RMSE(d_3), and RMSE(d_4) for unidimensional data
Figure 3a: RMSE(d_1)
Figure 3b: RMSE(d_2)
Figure 3c: RMSE(d_3)
Figure 3d: RMSE(d_4)

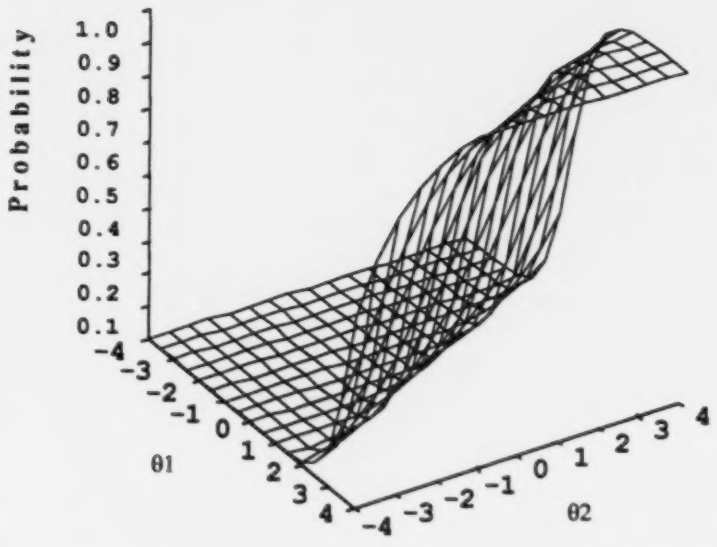
Figure 4. Bias(d_x)
Figure 4a: Bias(d_1)
Figure 4b: Bias(d_2)
Figure 4c: Bias(d_3)
Figure 4d: Bias(d_4)

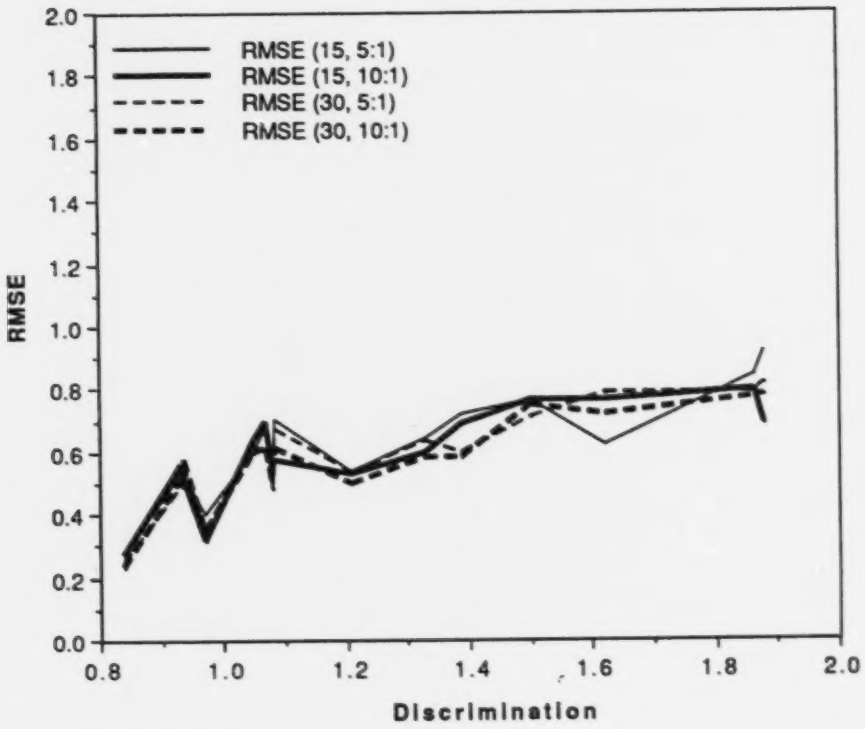
Figure 5. RMSE(θ) and Bias(θ) for unidimensional data
Figure 5a: RMSE(θ)
Figure 5b: Bias(θ)

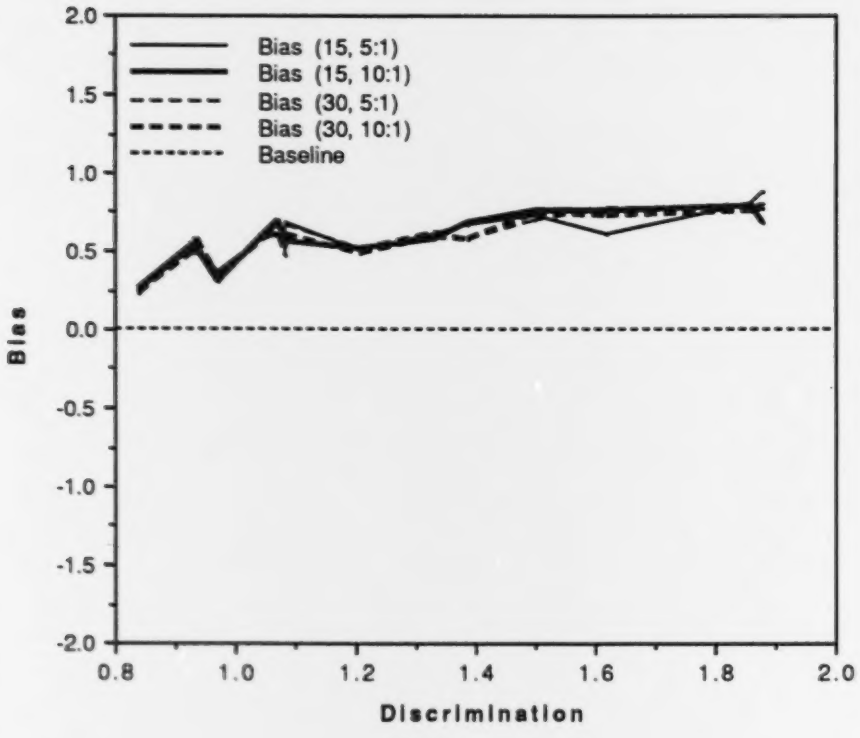
Figure 6. RMSE(a) for unequated unidimensional data

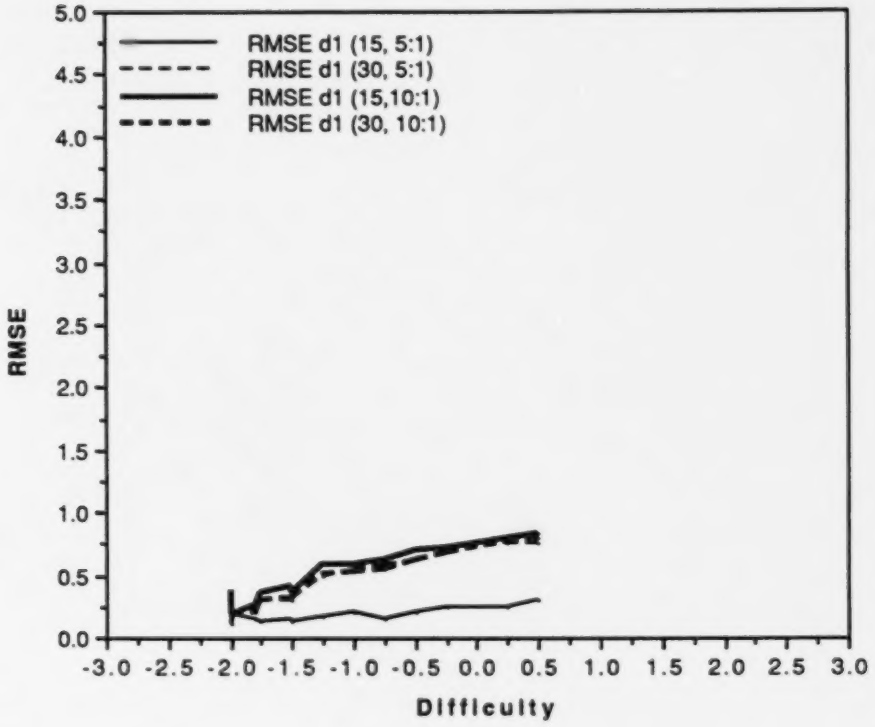

Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-4
-3
-2
-1
0
1
2
3
4
θ1
θ2

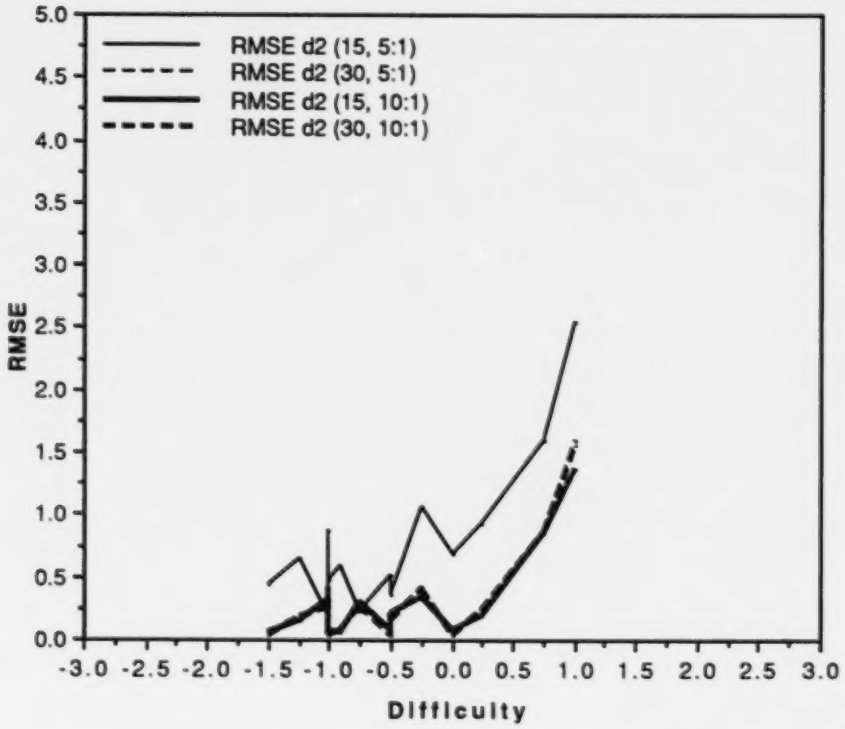

Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-4
-3
-2
-1
0
1
2
3
4
θ1
θ2

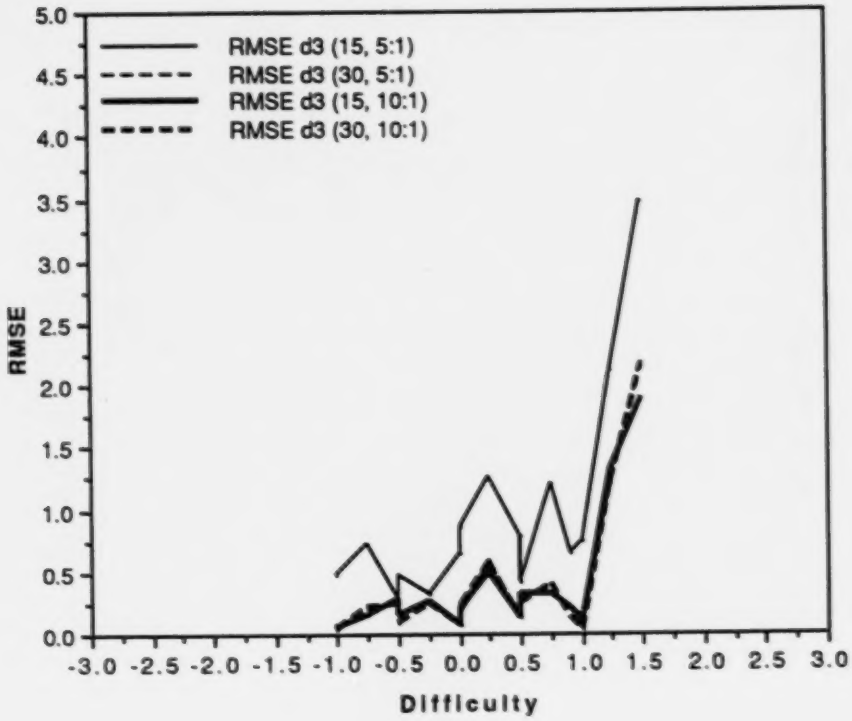

Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-4
-3
-2
-1
0
1
2
3
4
θ1
θ2

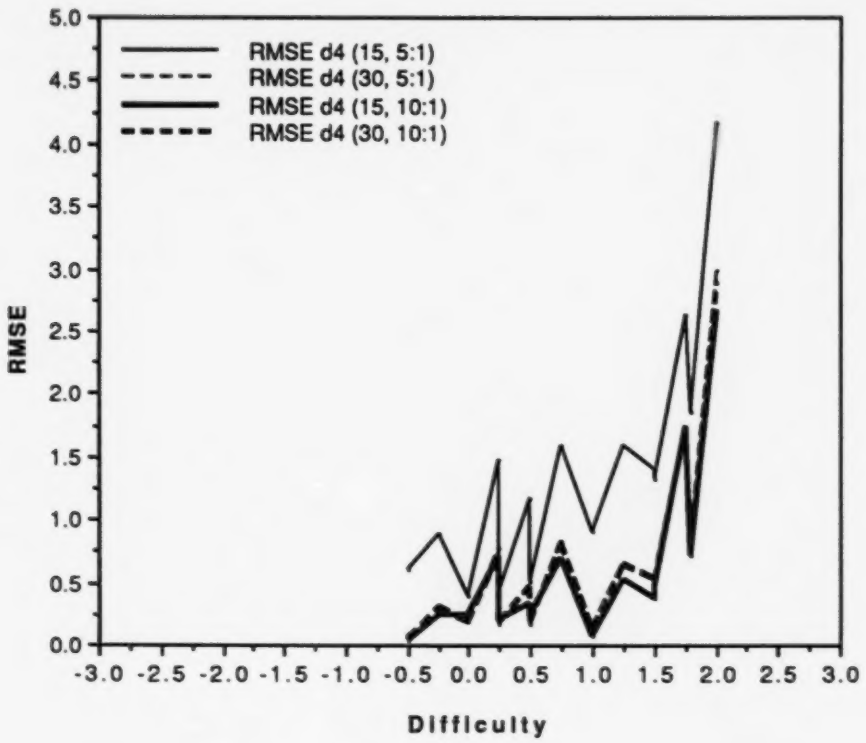

RMSE (15, 5:1)
RMSE (15, 10:1)
RMSE (30, 5:1)
RMSE (30, 10:1)
RMSE
Discrimination
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0

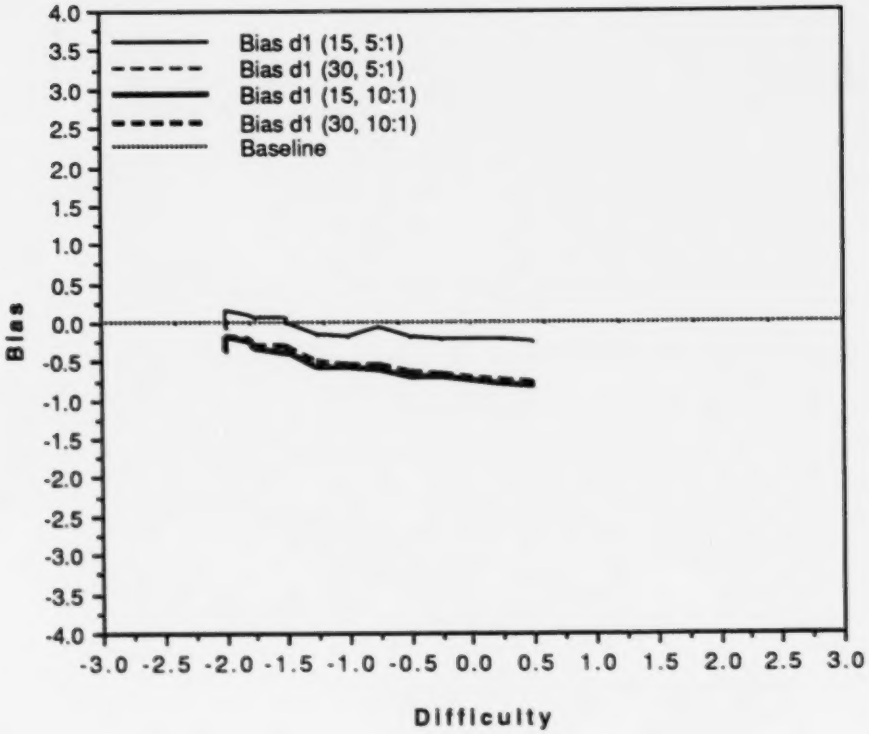

Bias (15, 5:1)
Bias (15, 10:1)
Bias (30, 5:1)
Bias (30, 10:1)
Baseline
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Bias
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Discrimination

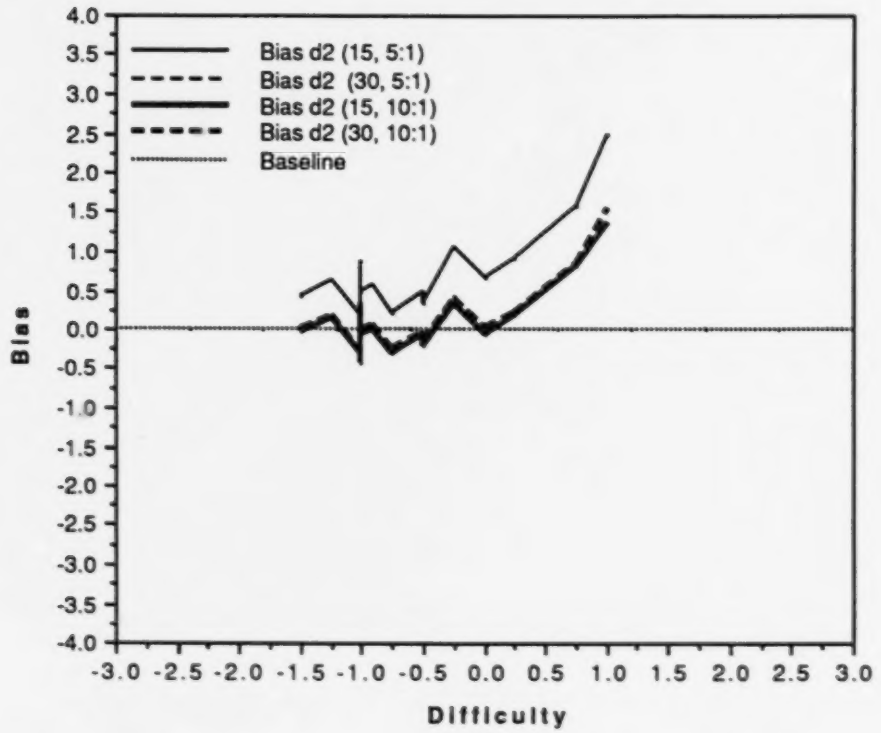

RMSE d1 (15, 5:1)
RMSE d1 (30, 5:1)
RMSE d1 (15,10:1)
RMSE d1 (30, 10:1)
RMSE
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Difficulty

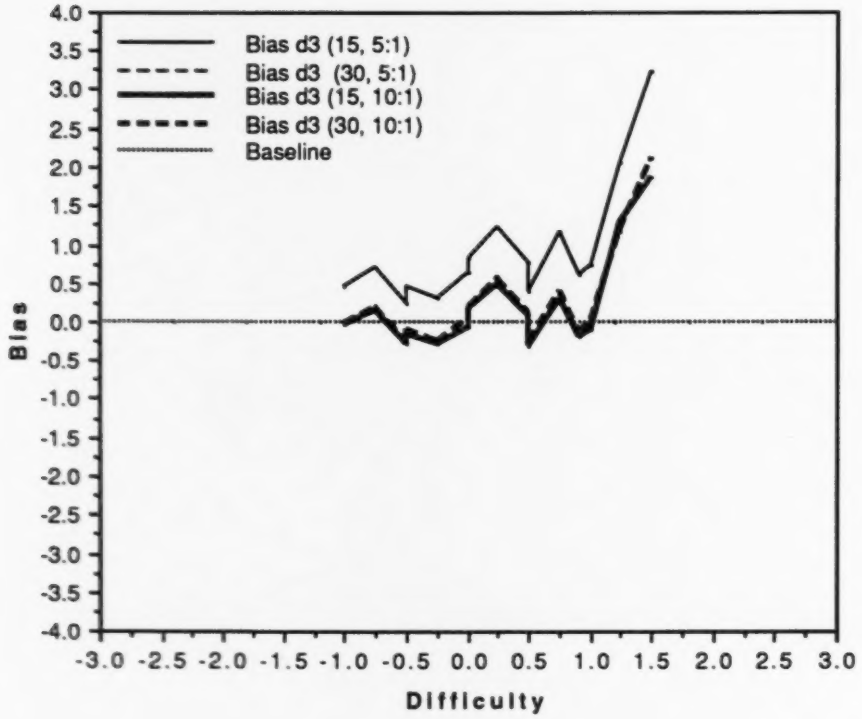

RMSE d2 (15, 5:1)
RMSE d2 (30, 5:1)
RMSE d2 (15, 10:1)
RMSE d2 (30, 10:1)
RMSE
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Difficulty

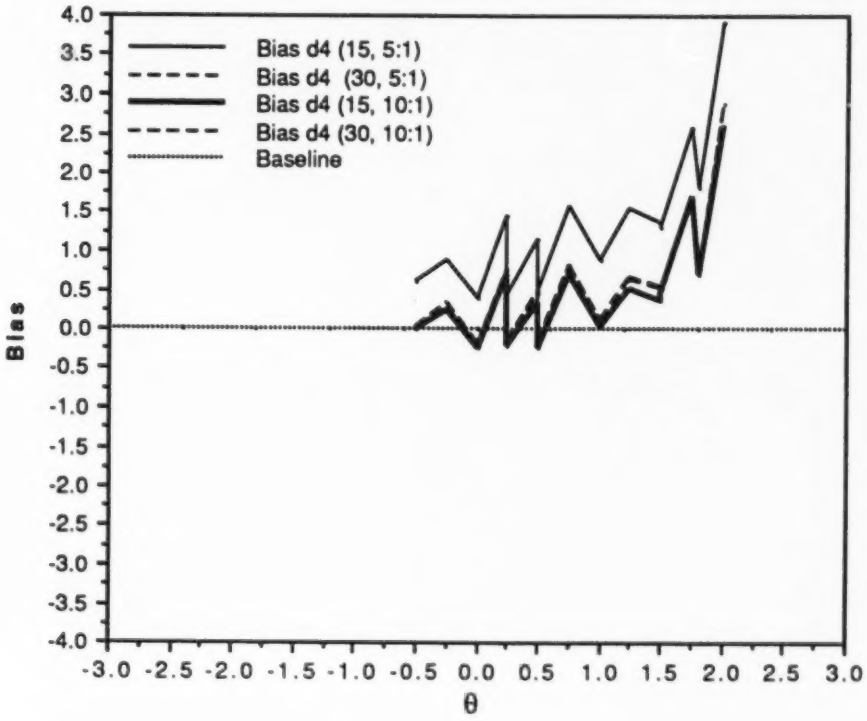

RMSE d3 (15, 5:1)
RMSE d3 (30, 5:1)
RMSE d3 (15, 10:1)
RMSE d3 (30, 10:1)
RMSE
Difficulty
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0

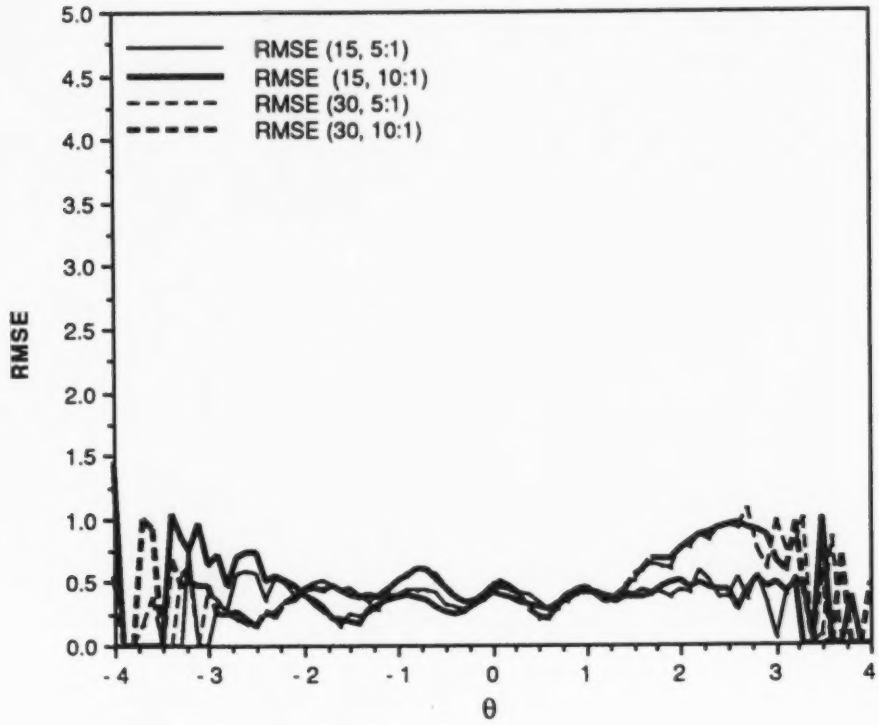

RMSE d4 (15, 5:1)
RMSE d4 (30, 5:1)
RMSE d4 (15, 10:1)
RMSE d4 (30, 10:1)
RMSE
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Difficulty

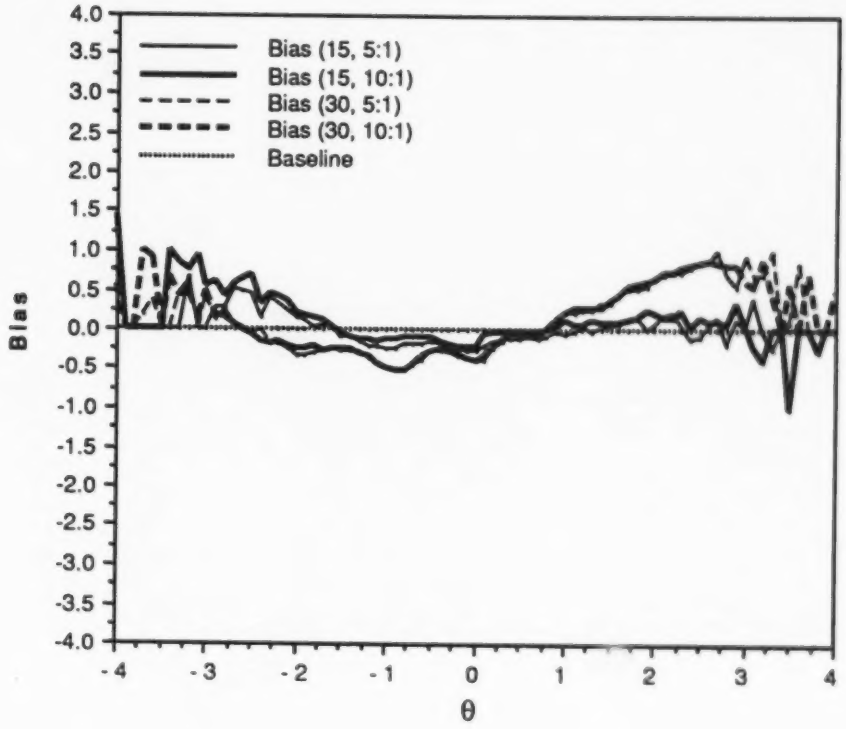

Bias d1 (15, 5:1)
Bias d1 (30, 5:1)
Bias d1 (15, 10:1)
Bias d1 (30, 10:1)
Baseline
Bias
Difficulty
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0

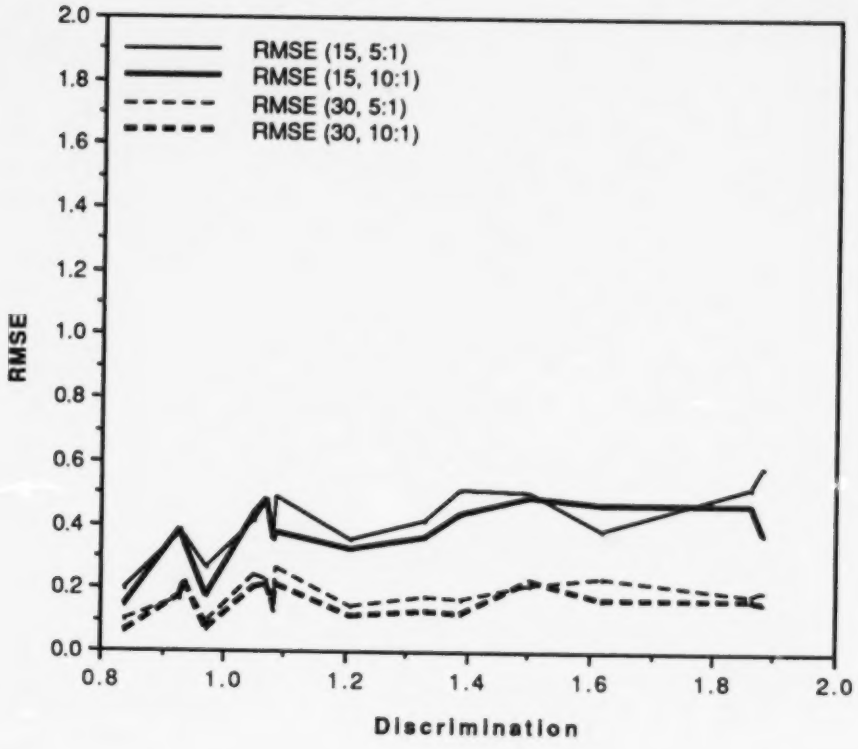

Bias d2 (15, 5:1)
Bias d2 (30, 5:1)
Bias d2 (15, 10:1)
Bias d2 (30, 10:1)
Baseline
Bias
Difficulty
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0


Bias d3 (15, 5:1)
Bias d3 (30, 5:1)
Bias d3 (15, 10:1)
Bias d3 (30, 10:1)
Baseline
Bias
Difficulty
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0


Bias d4 (15, 5:1)
Bias d4 (30, 5:1)
Bias d4 (15, 10:1)
Bias d4 (30, 10:1)
Baseline
Bias
θ
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0


RMSE (15, 5:1)
RMSE (15, 10:1)
RMSE (30, 5:1)
RMSE (30, 10:1)
RMSE
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-4
-3
-2
-1
0
1
2
3
4
θ


Bias (15, 5:1)
Bias (15, 10:1)
Bias (30, 5:1)
Bias (30, 10:1)
Baseline
Bias
θ
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
-4
-3
-2
-1
0
1
2
3
4


RMSE (15, 5:1)
RMSE (15, 10:1)
RMSE (30, 5:1)
RMSE (30, 10:1)
RMSE
Discrimination
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0